M121

IMPORTANT NOTE CONCERNING THE FOLLOWING PAGES

THE PAGES WHICH FOLLOW HAVE BEEN FILMED TWICE IN ORDER TO OBTAIN THE BEST REPRODUCTIVE QUALITY

USERS SHOULD CONSULT ALL THE PAGES REPRODUCED ON THE FICHE IN ORDER TO OBTAIN A COMPLETE READING OF THE TEXT.

REMARQUE IMPORTANTE CONCERNANT LES PAGES QUI SUIVENT

LES PAGES SUIVANTES ONT ÉTÉ REPRODUITES EN DOUBLE AFIN D'AMÉLIORER LA QUALITÉ DE REPRODUCTION

LES UTILISATEURS DOIVENT CONSULTER TOUTES LES PAGES REPRODUITES SUR LA FICHE AFIN D'OBTENIR LA LECTURE DU TEXTE INTÉGRAL

Micromedia
ProQuest

20 Victoria Street
Toronto, Ontario M5C 2N8
Tel.: (416) 362-5211
Toll Free: 1-800-387-2689
Fax: (416) 362-6161
Email: info@micromedia.on.ca

OQRE

Office de
la qualité et
de la responsabilité
en éducation

L'administration du Test provincial de compétences linguistiques d'octobre 2002 : Explication technique des modèles de mise en équivalence et du processus utilisé pour établir le seuil de réussite

L'administration du Test provincial de compétences linguistiques d'octobre 2002 : Explication technique des modèles de mise en équivalence et du processus utilisé pour établir le seuil de réussite

Introduction

Le présent document technique vient compléter le document intitulé « *L'administration du Test provincial de compétences linguistiques d'octobre 2002 : Résumé des modèles de mise en équivalence et du processus utilisé pour établir le seuil de réussite* », publié par l'Office de la qualité et de la responsabilité en éducation (OQRE). En plus des renseignements fournis dans le résumé, le présent document explique les processus utilisés pour relier les résultats du test administré en octobre 2002 (p. ex., plan détaillé du test, clés de réponses, seuil de réussite) à ceux du test administré en février 2002. Ces renseignements sont fournis en deux parties. La première partie décrit la composante lecture, et la deuxième traite de la composante écriture.

Analyse raisonnée de la mise en équivalence

Un processus de mise en équivalence est nécessaire quand les résultats d'une administration doivent être comparables et interchangeables avec les résultats d'une autre administration. La mise en équivalence compense les variations entre les niveaux de difficulté des deux tests, conçus pour être semblables en matière de contenu et de difficulté (Kolen & Brennan, 1999; Yang & Houang, 1996).

Première partie : la composante lecture

Plan détaillé du test

Les items du test pour la composante lecture sont tirés des domaines et des contenus d'apprentissage des programmes-cadres du *Curriculum de l'Ontario* pour toutes les matières jusqu'à la fin de la 9e année. Le document « Liens du TPCL avec le curriculum (2003) » récemment publié montre des exemples de liens entre le contenu du Test provincial de compétences linguistiques (TPCL) et *Le curriculum de l'Ontario.*

Le plan détaillé du test pour la composante lecture peut être décrit en termes des trois types de textes pour la composante lecture et des trois types d'habiletés, comme le montre le tableau 1 ci-dessous :

Tableau 1 : Plan détaillé pour la composante lecture par type d'habiletés

Type d'habileté	Types de textes de lecture de langue anglaise				Types de textes de lecture de langue française			
	Informatifs	Courants avec éléments graphiques	Narratifs	Total	Informatifs	Courants avec éléments graphiques	Narratifs	Total
Habileté 1	15	8	7	30	15	8	7	30
Habileté 2	20	11	14	45	20	11	14	45
Habileté 3	10	6	9	25	10	6	9	25
Total	45	25	30	100	45	25	30	100

Remarque : Habileté 1 = compréhension des idées et des renseignements directement énoncés (explicites); Habileté 2 = compréhension des idées et des renseignements indirectement énoncés (implicites); et Habileté 3 = établissement de liens entre les connaissances et expériences personnelles et les idées et renseignements figurant dans les textes (interprétation du sens et déduction).

La composante lecture comprend 40 questions à réponse choisie, 35 questions à réponse construite et 25 questions à réponse construite avec explication. Les questions à réponse choisie et les questions à réponse construite ont été notées de la façon suivante : 0 point pour une réponse incorrecte et 2 points pour une réponse correcte. Les questions à réponse construite avec explication ont été notées sur une échelle de trois points : 0 point pour une réponse incorrecte, 1 point pour une réponse partiellement correcte et 2 points pour une réponse correcte.

Élaboration du test

Des éducatrices et éducateurs expérimentés ayant une connaissance approfondie du *Curriculum de l'Ontario* ont mis au point les items pour la composante lecture. Tous les items de lecture utilisés durant le test administré en octobre 2002 ont été mis à l'essai en mars 2000 dans des écoles choisies dans différentes régions de la province, y compris des écoles offrant les cours théoriques et appliqués. (La participation des écoles était volontaire.) Les items ont ensuite été révisés au besoin et mis à l'essai une deuxième fois durant l'administration du test de février 2002, à l'aide d'un modèle en spirale. C'est-à-dire que durant l'administration du test de février 2002, des questions faisant partie de la mise à l'essai ont été intégrées au test actuel et attribuées à des élèves de toutes les régions de la province de façon systématique et aléatoire. Cette approche de mise à l'essai ressemble au modèle utilisé dans plusieurs autres territoires (Smisko, Twing, & Denny, 2000). L'utilisation du modèle en spirale a permis à l'OQRE de mettre à l'essai une quantité suffisante de matériel d'évaluation pour le test qui serait administré en octobre 2002, y compris le matériel d'appoint requis. Après la mise à l'essai, les items ont été analysés et choisis en fonction du plan détaillé du test ainsi que diverses exigences statistiques (p. ex., Crocker & Algina, 1986; voir *XCALIBRE User's Manual*).

Exigences psychométriques (« Lignes directrices ») pour la mise en équivalence

Un grand nombre de textes traitent des exigences psychométriques reliées à la mise en équivalence d'évaluations (p. ex., Crocker & Algina, 1986; Kolen & Brennan, 1999; Muraki, Himbo & Lee, 2000; Yang & Houang, 1996). Selon Crocker et Algina, le processus de mise en équivalence doit être choisi en fonction de trois critères principaux, soit « convenance des hypothèses, valeur concrète et exactitude ». (p. 464). Cependant, Kolen et Brennan (1999, pages 283-285) fournissent une liste plus complète de conditions (« lignes directrices ») permettant

Tableau 1 : Plan détaillé pour la composante lecture par type d'habiletés

Type d'habileté	Types de textes de lecture de langue anglaise				Types de textes de lecture de langue française			
	Informatifs	Courants avec éléments graphiques	Narratifs	Total	Informatifs	Courants avec éléments graphiques	Narratifs	Total
Habileté 1	15	8	7	30	15	8	7	30
Habileté 2	20	11	14	45	20	11	14	45
Habileté 3	10	6	9	25	10	6	9	25
Total	45	25	30	100	45	25	30	100

Remarque : Habileté 1 = compréhension des idées et des renseignements directement énoncés (explicites); Habileté 2 = compréhension des idées et des renseignements indirectement énoncés (implicites); et Habileté 3 = établissement de liens entre les connaissances et expériences personnelles et les idées et renseignements figurant dans les textes (interprétation du sens et déduction).

La composante lecture comprend 40 questions à réponse choisie, 35 questions à réponse construite et 25 questions à réponse construite avec explication. Les questions à réponse choisie et les questions à réponse construite ont été notées de la façon suivante : 0 point pour une réponse incorrecte et 2 points pour une réponse correcte. Les questions à réponse construite avec explication ont été notées sur une échelle de trois points : 0 point pour une réponse incorrecte, 1 point pour une réponse partiellement correcte et 2 points pour une réponse correcte.

Élaboration du test

Des éducatrices et éducateurs expérimentés ayant une connaissance approfondie du *Curriculum de l'Ontario* ont mis au point les items pour la composante lecture. Tous les items de lecture utilisés durant le test administré en octobre 2002 ont été mis à l'essai en mars 2000 dans des écoles choisies dans différentes régions de la province, y compris des écoles offrant les cours théoriques et appliqués. (La participation des écoles était volontaire.) Les items ont ensuite été révisés au besoin et mis à l'essai une deuxième fois durant l'administration du test de février 2002, à l'aide d'un modèle en spirale. C'est-à-dire que durant l'administration du test de février 2002, des questions faisant partie de la mise à l'essai ont été intégrées au test actuel et attribuées à des élèves de toutes les régions de la province de façon systématique et aléatoire. Cette approche de mise à l'essai ressemble au modèle utilisé dans plusieurs autres territoires (Smisko, Twing, & Denny, 2000). L'utilisation du modèle en spirale a permis à l'OQRE de mettre à l'essai une quantité suffisante de matériel d'évaluation pour le test qui serait administré en octobre 2002, y compris le matériel d'appoint requis. Après la mise à l'essai, les items ont été analysés et choisis en fonction du plan détaillé du test ainsi que diverses exigences statistiques (p. ex., Crocker & Algina, 1986; voir *XCALIBRE User's Manual*).

Exigences psychométriques (« Lignes directrices ») pour la mise en équivalence

Un grand nombre de textes traitent des exigences psychométriques reliées à la mise en équivalence d'évaluations (p. ex., Crocker & Algina, 1986; Kolen & Brennan, 1999; Muraki, Himbo & Lee, 2000; Yang & Houang, 1996). Selon Crocker et Algina, le processus de mise en équivalence doit être choisi en fonction de trois critères principaux, soit « convenance des hypothèses, valeur concrète et exactitude ». (p. 464). Cependant, Kolen et Brennan (1999, pages 283-285) fournissent une liste plus complète de conditions (« lignes directrices ») permettant

d'obtenir une mise en équivalence satisfaisante. Le Tableau 2 fournit une version abrégée de la liste de conditions de Kolen et Brennan et de la façon dont elles peuvent être utilisées pour placer dans un contexte le processus de mise en équivalence pour la composante lecture du test d'octobre 2002.

Tableau 2 : Exigences psychométriques (« Lignes directrices ») pour la mise en équivalence

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante lecture
Buts de la mise en équivalence	Les buts de la mise en équivalence doivent être clairement précisés, surtout en ce qui concerne l'exactitude de la mise en équivalence et le degré auquel les notes peuvent être comparées sur de longues périodes de temps.	L'objectif à long terme de l'OQRE est d'élaborer et d'administrer des instruments d'évaluation conformes aux attentes détaillées dans *Le curriculum de l'Ontario*. Les items pour chaque instrument d'évaluation sont élaborés conformément au plan détaillé du test.
Plan détaillé du test	Le contenu du test et les exigences statistiques doivent être bien définies et stables dans le temps.	Le contenu du test et les exigences statistiques sont identiques pour les tests administrés en février 2002 et en octobre 2002. Les mêmes paramètres de théorie classique et de la réponse à l'item ont été utilisés pour guider le choix et la révision des items.
Longueur du test	Le test doit être suffisamment long (au moins 30 items).	La composante lecture du TPCL comprend 100 items de lecture (40 questions à réponse choisie, 35 questions à réponse construite et 25 questions à réponse construite avec explication).
Clés de réponses	Les clés de réponses doivent être stables quand les items sont utilisés plus d'une fois.	Les clés de réponses ont été mises à l'essai en mars 2000 et validés au cours de la notation du matériel mis à l'essai en février/mars 2002.
Représentation du contenu	Le contenu et les caractéristiques statistiques de l'ensemble d'items communs doivent être représentatifs du test au complet.	Les items dont les tests de février 2002 et d'octobre 2002 sont composés proviennent de domaines ayant un contenu semblable et des contenus d'apprentissage tirés du *Curriculum de l'Ontario*.
Proportion d'items communs	Le nombre d'items communs doit être suffisamment élevé, au moins 20 p. 100 du test pour les tests comprenant 40 items ou plus.	Tous les textes de lecture du test d'octobre 2002 ont été mis à l'essai durant l'évaluation complète de février 2002, à l'aide d'un modèle en spirale. Chaque texte de lecture utilisé durant le test d'octobre 2002 a une contrepartie dans le test de février 2002. Les traits communs ont été opérationalisés en termes de congruence du contenu et des types d'items.
Emplacement	L'ensemble d'items communs doit être mis à peu près au même endroit dans l'ancien et le nouveau test.	La diffusion d'une partie du TPCL nécessite le remplacement de certains textes. Cela crée une limite du côté pratique. De plus, pour faire une mise à l'essai d'un test au complet, il faudrait consacrer plus de temps durant la journée d'enseignement et le coût du test augmenterait (p. ex., impression, notation).

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante lecture
Horaire	Liaison double : une version antérieure devrait être administrée au même moment de l'année que la version devant être mise en équivalence. Une ancienne version devrait être administrée au cours de l'année passée.	L'administration habituelle du TPCL aux élèves de 10e année se fait en octobre (premier semestre). Il a fallu faire exception à cette règle durant l'année scolaire 2001-2002 à cause d'une infraction à la confidentialité du test survenue en octobre 2001. La mise à l'essai a eu lieu en février 2002 – huit mois avant le test d'octobre 2002.
Administration du test	Le test et les items dont il est composé doivent être sécurisés. Le test doit être administré dans des conditions soigneusement contrôlées et normalisées, identiques chaque fois que le test est administré.	Le TPCL est administré conformément aux exigences détaillées dans le *Guide pour l'administration du Test provincial de compétences linguistiques – octobre 2001* et le *Guide pour l'administration du Test provincial de compétences linguistiques, octobre 2002.*

En évaluant le contexte pour les modèles de mise en équivalence utilisés pour la composante lecture du test d'octobre 2002, il faut tenter de répondre à deux questions clés : a) Faut-il remplir toutes ces conditions? et b) Dans quelle mesure l'élaboration et la construction de la composante lecture sont-elles conformes à ces lignes directrices?

Question a) : Faut-il remplir toutes ces conditions? Selon Kolen et Brennan (1999), « il n'est pas nécessaire de remplir toutes ces conditions pour obtenir une mise en équivalence satisfaisante » (p. 283). Traub (1994) partage ce point de vue, comme on le voit ci-dessous :

> « Cependant, dans plusieurs applications, il n'est peut-être même pas important d'essayer d'obtenir une réponse positive ou négative à la question de parallélisme d'un ensemble de tests. Nous pourrions plutôt décider, après avoir examiné l'évidence disponible – moyennes des tests, écarts-types des tests et covariances – que deux instruments ou plus sont suffisamment parallèles pour que nous en tirions des avantages tout comme s'ils étaient réellement parallèles… » (p. 54).

Question b) : Dans quelle mesure l'élaboration et la conception de la composante lecture sont-elles conformes aux lignes directrices fournies par Kolen et Brennan? Il est clair que les tests de février 2002 et d'octobre 2002 sont équivalents et que les notes découlant de ces deux tests peuvent être comparées, pour les raisons suivantes :

1) les tests de février 2002 et d'octobre 2002 ont été élaborés en fonction du même plan détaillé de test, qui est basé sur *Le curriculum de l'Ontario* (congruence du contenu du test);
2) les tests de février 2002 et d'octobre 2002 ont été mis à l'essai pour la première fois en mars 2000, avec la même cohorte d'élèves provenant de différentes régions de la province;
3) les clés de réponses utilisées pour le test d'octobre 2002 ont été élaborées en mars 2000 et validées en février 2002, en même temps que les clés de réponses pour le test administré en février 2002;
4) tous les textes inclus dans la composante lecture du test d'octobre 2002 ont été intégrés au test administré en février 2002. Les items en lecture ont été analysés avec les items en lecture faisant

d'obtenir une mise en équivalence satisfaisante. Le Tableau 2 fournit une version abrégée de la liste de conditions de Kolen et Brennan et de la façon dont elles peuvent être utilisées pour placer dans un contexte le processus de mise en équivalence pour la composante lecture du test d'octobre 2002.

Tableau 2 : Exigences psychométriques (« Lignes directrices ») pour la mise en équivalence

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante lecture
Buts de la mise en équivalence	Les buts de la mise en équivalence doivent être clairement précisés, surtout en ce qui concerne l'exactitude de la mise en équivalence et le degré auquel les notes peuvent être comparées sur de longues périodes de temps.	L'objectif à long terme de l'OQRE est d'élaborer et d'administrer des instruments d'évaluation conformes aux attentes détaillées dans *Le curriculum de l'Ontario*. Les items pour chaque instrument d'évaluation sont élaborés conformément au plan détaillé du test.
Plan détaillé du test	Le contenu du test et les exigences statistiques doivent être bien définies et stables dans le temps.	Le contenu du test et les exigences statistiques sont identiques pour les tests administrés en février 2002 et en octobre 2002. Les mêmes paramètres de théorie classique et de la réponse à l'item ont été utilisés pour guider le choix et la révision des items.
Longueur du test	Le test doit être suffisamment long (au moins 30 items).	La composante lecture du TPCL comprend 100 items de lecture (40 questions à réponse choisie, 35 questions à réponse construite et 25 questions à réponse construite avec explication).
Clés de réponses	Les clés de réponses doivent être stables quand les items sont utilisés plus d'une fois.	Les clés de réponses ont été mises à l'essai en mars 2000 et validés au cours de la notation du matériel mis à l'essai en février/mars 2002.
Représentation du contenu	Le contenu et les caractéristiques statistiques de l'ensemble d'items communs doivent être représentatifs du test au complet.	Les items dont les tests de février 2002 et d'octobre 2002 sont composés proviennent de domaines ayant un contenu semblable et des contenus d'apprentissage tirés du *Curriculum de l'Ontario*.
Proportion d'items communs	Le nombre d'items communs doit être suffisamment élevé, au moins 20 p. 100 du test pour les tests comprenant 40 items ou plus.	Tous les textes de lecture du test d'octobre 2002 ont été mis à l'essai durant l'évaluation complète de février 2002, à l'aide d'un modèle en spirale. Chaque texte de lecture utilisé durant le test d'octobre 2002 a une contrepartie dans le test de février 2002. Les traits communs ont été opérationalisés en termes de congruence du contenu et des types d'items.
Emplacement	L'ensemble d'items communs doit être mis à peu près au même endroit dans l'ancien et le nouveau test.	La diffusion d'une partie du TPCL nécessite le remplacement de certains textes. Cela crée une limite du côté pratique. De plus, pour faire une mise à l'essai d'un test au complet, il faudrait consacrer plus de temps durant la journée d'enseignement et le coût du test augmenterait (p. ex., impression, notation).

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante lecture
Horaire	Liaison double : une version antérieure devrait être administrée au même moment de l'année que la version devant être mise en équivalence. Une ancienne version devrait être administrée au cours de l'année passée.	L'administration habituelle du TPCL aux élèves de 10e année se fait en octobre (premier semestre). Il a fallu faire exception à cette règle durant l'année scolaire 2001-2002 à cause d'une infraction à la confidentialité du test survenue en octobre 2001. La mise à l'essai a eu lieu en février 2002 – huit mois avant le test d'octobre 2002.
Administration du test	Le test et les items dont il est composé doivent être sécurisés. Le test doit être administré dans des conditions soigneusement contrôlées et normalisées, identiques chaque fois que le test est administré.	Le TPCL est administré conformément aux exigences détaillées dans le *Guide pour l'administration du Test provincial de compétences linguistiques – octobre 2001* et le *Guide pour l'administration du Test provincial de compétences linguistiques, octobre 2002.*

En évaluant le contexte pour les modèles de mise en équivalence utilisés pour la composante lecture du test d'octobre 2002, il faut tenter de répondre à deux questions clés : a) Faut-il remplir toutes ces conditions? et b) Dans quelle mesure l'élaboration et la construction de la composante lecture sont-elles conformes à ces lignes directrices?

Question a) : Faut-il remplir toutes ces conditions? Selon Kolen et Brennan (1999), « il n'est pas nécessaire de remplir toutes ces conditions pour obtenir une mise en équivalence satisfaisante » (p. 283). Traub (1994) partage ce point de vue, comme on le voit ci-dessous :

> « Cependant, dans plusieurs applications, il n'est peut-être même pas important d'essayer d'obtenir une réponse positive ou négative à la question de parallélisme d'un ensemble de tests. Nous pourrions plutôt décider, après avoir examiné l'évidence disponible – moyennes des tests, écarts-types des tests et covariances – que deux instruments ou plus sont suffisamment parallèles pour que nous en tirions des avantages tout comme s'ils étaient réellement parallèles… » (p. 54).

Question b) : Dans quelle mesure l'élaboration et la conception de la composante lecture sont-elles conformes aux lignes directrices fournies par Kolen et Brennan? Il est clair que les tests de février 2002 et d'octobre 2002 sont équivalents et que les notes découlant de ces deux tests peuvent être comparées, pour les raisons suivantes :

1) les tests de février 2002 et d'octobre 2002 ont été élaborés en fonction du même plan détaillé de test, qui est basé sur *Le curriculum de l'Ontario* (congruence du contenu du test);
2) les tests de février 2002 et d'octobre 2002 ont été mis à l'essai pour la première fois en mars 2000, avec la même cohorte d'élèves provenant de différentes régions de la province;
3) les clés de réponses utilisées pour le test d'octobre 2002 ont été élaborées en mars 2000 et validées en février 2002, en même temps que les clés de réponses pour le test administré en février 2002;
4) tous les textes inclus dans la composante lecture du test d'octobre 2002 ont été intégrés au test administré en février 2002. Les items en lecture ont été analysés avec les items en lecture faisant

partie de l'évaluation complète de février 2002 et les mêmes exigences statistiques ont été utilisées pour choisir ces items;

5) les items des tests de février 2002 et d'octobre 2002 ont été calibrés à l'aide de l'analyse selon la théorie de réponse à l'item;
6) les deux tests ont été administrés conformément à des processus normalisés et dans les mêmes conditions sécuritaires.

Modèles de mise en équivalence pour la composante lecture

Les méthodes de mise en équivalence utilisées pour obtenir des seuils de réussite équivalents pour les tests d'octobre et de février 2002 peuvent être groupées en trois catégories : 1) analyse selon la théorie classique d'évaluation, 2) analyse selon la théorie de la réponse à l'item (TRI) et 3) analyse de reregroupement. Chaque méthode est décrite ci-dessous.

1) Analyse selon la théorie classique d'évaluation

Trois techniques ont été utilisées, soit celle d'équation linéaire, d'équicentilage et d'équation de la moyenne.

Équation linéaire. On considère que les notes obtenues à deux tests ont été mises en équivalence si elles contiennent le même nombre d'écarts-types se situant au-dessus ou en-dessous de la moyenne pour un groupe d'élèves ayant fait le test. On traite ce sujet de manière détaillée dans les documents portant sur les évaluations (p. ex., Crocker & Algina, 1986; Kolen & Brennan, 1999; Petersen, Kolen & Hoover, 1989). Voir, par exemple, l'extrait ci-dessous concernant la définition, la formule et le processus de calcul :

> **Définition** : « En équation linéaire, une transformation est choisie de façon à ce que les notes aux deux tests soient considérées comme ayant été mises en équivalence si elles correspondent au même nombre d'écarts-types au-dessus (ou au-dessous) de la moyenne pour un groupe donné d'élèves faisant le test ». (Angoff [1994] cité dans Petersen, Kolen & Hoover, 1989, p. 251).
>
> **Formule** : « $(x - \mu_x)/\delta_x = (y - \mu_y)/\delta_y$
> Quand x et y se rapportent aux notes des tests devant être mis en équivalence et quand μ et δ se rapportent, respectivement, aux moyennes et aux écarts-types de x et y pour une population donnée d'élèves faisant les tests. » (p. 251).
>
> **Calcul** :
> a) Calculer la moyenne échantillon (m) et le ou les écarts-types échantillon (s) des notes pour les deux tests devant être mis en équivalence.
> b) Obtenir les notes mises en équivalence pour les deux tests en substituant ces valeurs à celles qui se trouvent dans l'équation précédente (p. 252).

Méthode d'équicentilage. Cette méthode détermine quelles populations d'élèves se sont classées au même rang-centile aux deux tests. Une discussion détaillée de cette technique est fournie par Kolen et Brennan (1999). Voir l'extrait ci-dessous décrivant la façon dont cette technique est opérationalisée.

> « Un rang-centile équivalent pour une population d'élèves peut être obtenu de la façon suivante : pour la note d'un test X donné, trouver le pourcentage d'élèves faisant le test en obtenant des notes équivalentes ou inférieures à cette note de test X. Ensuite, trouver la note du test Y ayant le même pourcentage d'élèves au même niveau ou à un niveau inférieur. Ces notes de test X et de test Y sont considérées comme étant équivalentes. » (Kolen et Brennan, 1999, pp. 36-37).

Mise en équivalence de la moyenne. Les notes du second test sont converties de façon à ce que la moyenne soit équivalente à la moyenne du premier test; la différence entre les moyennes représente la constante dans l'échelle de notes. (Barnard, 1996).

2) Analyse selon la théorie de la réponse à l'item (TRI)
Les items des tests de février 2002 et d'octobre 2002 ont été calibrés ensemble à l'aide des réponses de l'administration de février. Des estimations des habiletés des élèves ont été dérivées pour les élèves ayant fait le test en février. À l'aide des estimations des habiletés et des résultats obtenus par les élèves, nous avons établi quelles notes les mêmes élèves obtiendraient s'ils faisaient le test d'octobre 2002. De cette façon, les notes TRI des élèves pour les tests de février 2002 et d'octobre 2002 ont été placées sur la même échelle et ont pu être directement comparées.

Voici une description des étapes du processus de mise en équivalence TRI :

Première étape : calibrage des items TRI
Deux approches de notation ont été utilisées, soit une pour les items notés de façon dichotomique (p. ex., 0 = incorrect, 2 = correct) et une autre pour les items notés de façon polytomique à l'aide de trois catégories (p. ex., 0 = incorrect, 1 = partiellement correct, 2 = correct). Les items notés de façon dichotomique ont été calibrés à l'aide d'un modèle logistique à un paramètre qui établit le paramètre de pseudo-chance à 0 et force le paramètre de discrimination à être équivalent pour tous les items. Les items notés de façon polytomique ont été calibrés à l'aide du modèle de crédit partiel généralisé (Bock, 1972).

Deuxième étape : échelle de transformation linéaire des habiletés (Base TRI)
Comme le test d'octobre 2002 a été créé avec des items qui avaient été mis à l'essai en même temps que le test administré en février 2002, les items des deux tests ont été calibrés à l'aide des réponses des élèves qui ont fait le test en février 2002. Des estimations des habiletés des élèves ont été obtenues durant le calibrage des items. L'estimation des habiletés pour les élèves ayant fait le test de février 2002 a ensuite été utilisée pour obtenir une estimation de la note que l'on s'attendrait à ce que les élèves aient obtenue s'ils avaient fait le test d'octobre 2002. Ce processus de calibrage d'items conjoints élimine le besoin d'effectuer une transformation linéaire de l'échelle des habiletés.
Troisième étape : conversion des notes obtenues en octobre 2002 en notes applicables à l'administration de février 2002

Après avoir obtenu les paramètres des items pour tous les élèves dans chaque groupe de langue (soit un pour le français et un autre pour l'anglais), le rendement de chaque élève au test administré en février 2002 a été noté. Les notes TRI des élèves montrent ainsi le nombre de questions auxquelles l'élève a répondu, l'exactitude de ses réponses et les caractéristiques des items.

Les deux courbes caractéristiques du test (CCT; soit, une pour le test de février 2002 et l'autre pour le test d'octobre 2002) ont été mises en correspondance sur la même échelle des habiletés et converties à nouveau en chiffres précis.

Les figures 1 et 2 montrent les CCT respectivement pour le français et l'anglais :

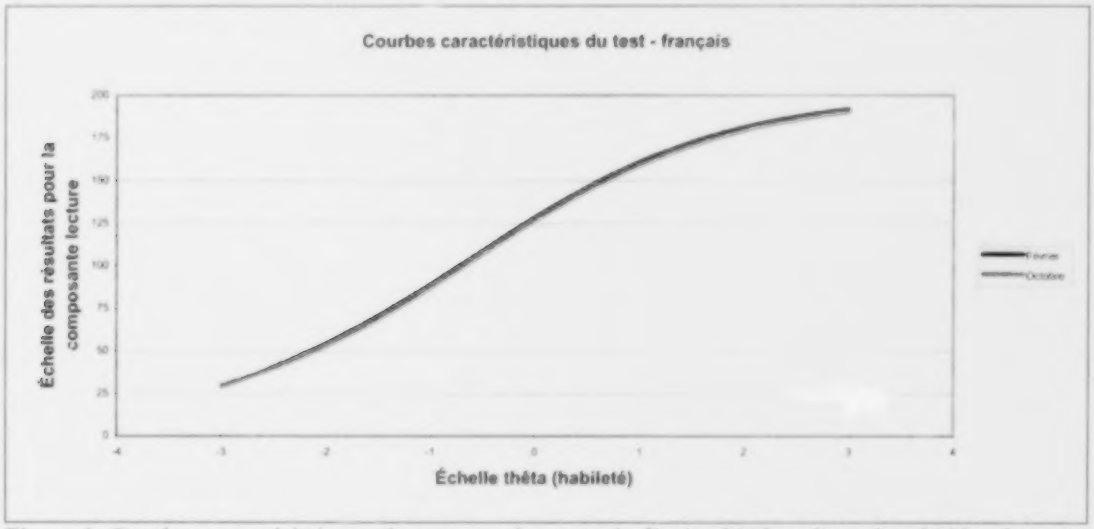

Figure 1. Courbes caractéristiques du test pour les tests de février 2002 et d'octobre 2002 – français.

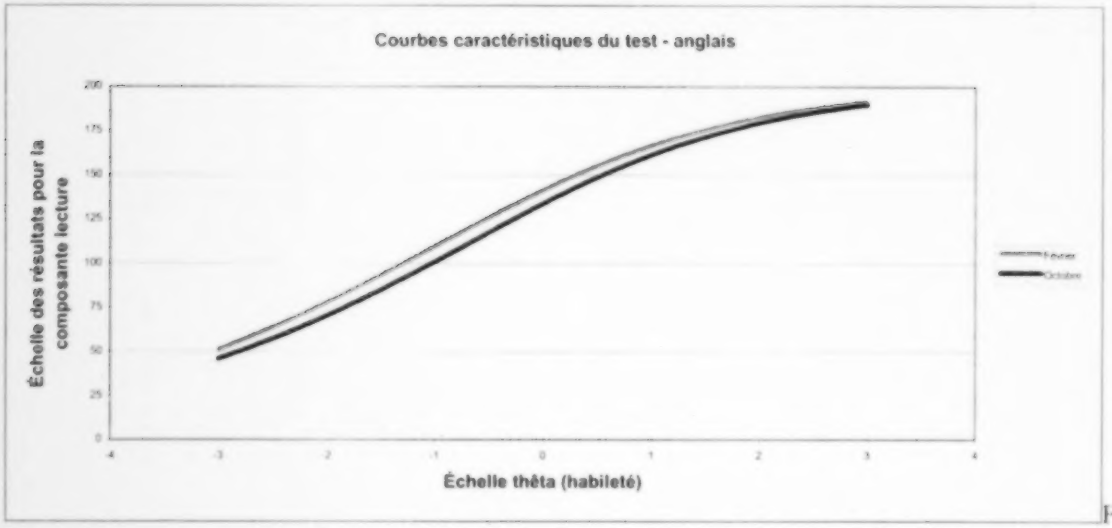

Figure 2. Courbes caractéristiques du test pour les tests de février 2002 et d'octobre 2002 – anglais.

3) Analyse de regroupement

Une analyse de regroupement est en même temps une technique exploratoire et confirmative. L'analyse de regroupement peut être utilisée pour une analyse exploratoire quand le nombre de points représentant le seuil de réussite n'est pas connu. La technique peut être utilisée comme analyse confirmative quand les notes qui représentent le seuil de réussite sont connues et doivent être validées (Sireci, 2001).

L'extrait ci-dessous donne une définition de l'analyse de regroupement et de son application à l'analyse de la notation du test.

> « [L'objectif d'une analyse de regroupement] est de répartir des cas (personnes, choses, événements, etc.) en groupes, ou types, de façon à ce que le degré d'association soit solide entre les membres du même groupe et faible entre les membres de groupes différents. Chaque groupe décrit donc, en matière des données recueillies, la catégorie à laquelle ses membres appartiennent. »
> (Source : http://www.clustan.com/what_is_cluster_analysis.html)
>
> « Si on considère que les personnes qui font des tests sont des « objets » devant être groupés, l'objectif de l'analyse de regroupement est de les grouper en catégories de façon à ce que les personnes faisant les tests à l'intérieur de chaque catégorie se ressemblent, mais soient différentes des personnes faisant des tests groupées dans d'autres catégories. » (Sireci, 2001, p. 341.)

Pour l'administration d'octobre 2002, l'analyse de regroupement incluait tous les types d'items (questions à réponse choisie, à réponse construite et à réponse construite avec explication), a été utilisée pour grouper les élèves en catégories selon leur rendement. Le long du continuum de rendement, un seuil de réussite a été établi afin de séparer les élèves qui avaient réussi de ceux qui n'avaient pas réussi.

Résultats – seuil de réussite final pour la composante lecture
À l'aide des techniques d'analyse CTT et de regroupement, une échelle de seuil de réussite a été établie. Cette échelle s'étendait de 118 à 124 pour les élèves de langue française et de 117 à 120 pour les élèves de langue anglaise. L'analyse TRI a été utilisée pour déterminer le seuil de réussite final, soit 118 sur un maximum de 200 points pour les élèves de langue française et 119 sur un maximum de 200 points pour les élèves de langue anglaise. Ces seuils de réussite sont équivalents aux seuils de réussite de février 2002. Bien qu'il soit possible de trouver des avantages et des inconvénients aux diverses méthodes de mise en équivalence (Kolen & Brennan, 1999), l'analyse TRI est avantageuse parce qu'elle permet de calibrer en même temps des questions de différentes versions. Les deux séries de notes sont placées sur la même échelle, ce qui permet de comparer directement les résultats.

Deuxième partie : la composante écriture

Les tâches d'écriture sont tirées des attentes et des contenus d'apprentissage des programmes-cadres du *Curriculum de l'Ontario* pour toutes les matières jusqu'à la fin de la 9e année. Le document « Liens du TPCL avec le curriculum (2003) » récemment publié montre des exemples de liens entre le contenu du TPCL et *Le curriculum de l'Ontario*.

Comme pour la composante lecture du TPCL, la liste de conditions de Kolen et Brennan est appliquée afin d'établir un contexte pour la comparaison entre les composantes écriture des tests de février 2002 et d'octobre 2002. Il faut réfléchir à deux questions clés : a) De quelle façon les tâches d'écriture des tests de février 2002 et d'octobre 2002 sont-elles comparables? et b) Une mise en équivalence peut-elle être établie pour les notes obtenues des deux ensembles de tâches d'écriture?

Tableau 3 : Exigences psychométriques (« Lignes directrices ») pour la mise en équivalence

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante écriture
Buts de la mise en équivalence	Les buts de la mise en équivalence doivent être clairement précisés, surtout en ce qui concerne l'exactitude de la mise en équivalence et le degré auquel les notes peuvent être comparées sur de longues périodes de temps.	L'objectif à long terme de l'OQRE est d'élaborer et d'administrer des instruments d'évaluation conformes aux attentes détaillées dans *Le curriculum de l'Ontario*. Les items pour chaque instrument d'évaluation sont élaborés conformément au plan détaillé du test.
Plan détaillé du test	Le contenu du test doit être bien défini et stable dans le temps.	Pour les tests de février 2002 et d'octobre 2002, la composante écriture comprend quatre tâches du même genre : • résumé • texte d'opinion en quelques paragraphes • nouvelle journalistique • texte informatif
Longueur du test	La longueur du test doit être standardisée.	Aucun changement entre le test de février 2002 et celui d'octobre 2002 :

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante écriture
		• résumé – moins de 100 mots • texte d'opinion – au moins trois paragraphes; trois pages lignées fournies, indiquant la longueur approximative requise • nouvelle journalistique – une page lignée à simple interligne fournie, indiquant la longueur approximative requise • texte informatif – espace ligné fourni, représentant environ une page
Grilles de notation	Les grilles de notation doivent être stables.	Aucun changement dans les grilles utilisées pour la notation selon l'approche globale entre le test de février 2002 et celui d'octobre 2002.
Représentation du contenu	Les caractéristiques du contenu des tâches d'écriture doivent être représentatives de la composante écriture du test.	Les critères de représentation du contenu doivent être comparables entre les tests de février 2002 et d'octobre 2002. • contenu canadien • sujets traitant des élèves ou de l'école • destinataires : adultes pour trois tâches (résumé, texte d'opinion, texte informatif) • texte informatif : la réponse de l'élève doit inclure au moins quatre points à l'appui choisis parmi un total de 18 points
Emplacement	Les tâches d'écriture doivent être placées au même endroit dans les anciens et les nouveaux tests.	Les tâches d'écriture ont été placées au même endroit dans les tests d'octobre 2002 et de février 2002.
Horaire	Liaison double : une version antérieure devrait être administrée au même moment de l'année que la version devant être mise en équivalence. Une version antérieure devrait être administrée au cours de l'année passée.	L'administration habituelle du TPCL aux élèves de 10e année se fait en octobre (premier semestre). Il a fallu faire exception à cette règle durant l'année scolaire 2001-2002 à cause d'une infraction à la confidentialité du test survenue en octobre 2001. La mise à l'essai a eu lieu en février 2002 – huit mois avant le test d'octobre 2002.
Administration du test	La confidentialité du test doit être assurée. Le test doit être administré dans des conditions soigneusement contrôlées et normalisées, identiques chaque fois que le test est administré.	Le TPCL est administré conformément aux exigences détaillées dans le *Guide pour l'administration du Test provincial de compétences linguistiques – octobre 2001* et le *Guide pour l'administration du Test provincial de compétences linguistiques, octobre 2002*. Aucun changement n'a été apporté au processus d'administration en matière du nombre de jours ou de la séquence des cahiers et des tâches. Chaque administration du test dure deux demi-journées.

Résultats – seuil de réussite final pour la composante lecture
À l'aide des techniques d'analyse CTT et de regroupement, une échelle de seuil de réussite a été établie. Cette échelle s'étendait de 118 à 124 pour les élèves de langue française et de 117 à 120 pour les élèves de langue anglaise. L'analyse TRI a été utilisée pour déterminer le seuil de réussite final, soit 118 sur un maximum de 200 points pour les élèves de langue française et 119 sur un maximum de 200 points pour les élèves de langue anglaise. Ces seuils de réussite sont équivalents aux seuils de réussite de février 2002. Bien qu'il soit possible de trouver des avantages et des inconvénients aux diverses méthodes de mise en équivalence (Kolen & Brennan, 1999), l'analyse TRI est avantageuse parce qu'elle permet de calibrer en même temps des questions de différentes versions. Les deux séries de notes sont placées sur la même échelle, ce qui permet de comparer directement les résultats.

Deuxième partie : la composante écriture

Les tâches d'écriture sont tirées des attentes et des contenus d'apprentissage des programmes-cadres du *Curriculum de l'Ontario* pour toutes les matières jusqu'à la fin de la 9e année. Le document « Liens du TPCL avec le curriculum (2003) » récemment publié montre des exemples de liens entre le contenu du TPCL et *Le curriculum de l'Ontario*.

Comme pour la composante lecture du TPCL, la liste de conditions de Kolen et Brennan est appliquée afin d'établir un contexte pour la comparaison entre les composantes écriture des tests de février 2002 et d'octobre 2002. Il faut réfléchir à deux questions clés : a) De quelle façon les tâches d'écriture des tests de février 2002 et d'octobre 2002 sont-elles comparables? et b) Une mise en équivalence peut-elle être établie pour les notes obtenues des deux ensembles de tâches d'écriture?

Tableau 3 : Exigences psychométriques (« Lignes directrices ») pour la mise en équivalence

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante écriture
Buts de la mise en équivalence	Les buts de la mise en équivalence doivent être clairement précisés, surtout en ce qui concerne l'exactitude de la mise en équivalence et le degré auquel les notes peuvent être comparées sur de longues périodes de temps.	L'objectif à long terme de l'OQRE est d'élaborer et d'administrer des instruments d'évaluation conformes aux attentes détaillées dans *Le curriculum de l'Ontario*. Les items pour chaque instrument d'évaluation sont élaborés conformément au plan détaillé du test.
Plan détaillé du test	Le contenu du test doit être bien défini et stable dans le temps.	Pour les tests de février 2002 et d'octobre 2002, la composante écriture comprend quatre tâches du même genre : • résumé • texte d'opinion en quelques paragraphes • nouvelle journalistique • texte informatif
Longueur du test	La longueur du test doit être standardisée.	Aucun changement entre le test de février 2002 et celui d'octobre 2002 :

Exigence	Description des exigences (source : Kolen & Brennan, 1999)	TPCL, octobre 2002, composante écriture
		• résumé – moins de 100 mots • texte d'opinion – au moins trois paragraphes; trois pages lignées fournies, indiquant la longueur approximative requise • nouvelle journalistique – une page lignée à simple interligne fournie, indiquant la longueur approximative requise • texte informatif – espace ligné fourni, représentant environ une page
Grilles de notation	Les grilles de notation doivent être stables.	Aucun changement dans les grilles utilisées pour la notation selon l'approche globale entre le test de février 2002 et celui d'octobre 2002
Représentation du contenu	Les caractéristiques du contenu des tâches d'écriture doivent être représentatives de la composante écriture du test.	Les critères de représentation du contenu doivent être comparables entre les tests de février 2002 et d'octobre 2002. • contenu canadien • sujets traitant des élèves ou de l'école • destinataires : adultes pour trois tâches (résumé, texte d'opinion, texte informatif) • texte informatif : la réponse de l'élève doit inclure au moins quatre points à l'appui choisis parmi un total de 18 points
Emplacement	Les tâches d'écriture doivent être placées au même endroit dans les anciens et les nouveaux tests.	Les tâches d'écriture ont été placées au même endroit dans les tests d'octobre 2002 et de février 2002.
Horaire	Liaison double : une version antérieure devrait être administrée au même moment de l'année que la version devant être mise en équivalence. Une version antérieure devrait être administrée au cours de l'année passée.	L'administration habituelle du TPCL aux élèves de 10e année se fait en octobre (premier semestre). Il a fallu faire exception à cette règle durant l'année scolaire 2001-2002 à cause d'une infraction à la confidentialité du test survenue en octobre 2001. La mise à l'essai a eu lieu en février 2002 – huit mois avant le test d'octobre 2002.
Administration du test	La confidentialité du test doit être assurée. Le test doit être administré dans des conditions soigneusement contrôlées et normalisées, identiques chaque fois que le test est administré.	Le TPCL est administré conformément aux exigences détaillées dans le *Guide pour l'administration du Test provincial de compétences linguistiques – octobre 2001* et le *Guide pour l'administration du Test provincial de compétences linguistiques, octobre 2002.* Aucun changement n'a été apporté au processus d'administration en matière du nombre de jours ou de la séquence des cahiers et des tâches. Chaque administration du test dure deux demi-journées

Établissement du seuil de réussite et résultats en écriture
La méthodologie d'établissement des normes connue sous le nom de « la méthodologie de classification des cahiers » a été utilisée pour obtenir des seuils de réussite équivalents pour les tests d'octobre 2002 et de février 2002. Ce processus nécessite de travailler avec des travaux d'élèves corrects et incorrects tirés de l'évaluation de février 2002, puis de faire une mise en équivalence entre ce matériel et du matériel noté, tiré de l'évaluation d'octobre 2002. Le processus de mise en équivalence a été basé sur un processus d'établissement des normes établi en novembre 2002.

Voici le processus de notation qui a été utilisé pour le test d'octobre 2002 :

Processus de notation des travaux en écriture. La composante écriture a été notée en deux phases.

Notation selon l'approche globale, Phase 1. Le travail de chaque élève a été noté selon l'approche globale. Donc, à mesure que chaque tâche a été notée, toutes les caractéristiques de la tâche ont été évaluées en même temps, puis notées. Une autre correctrice ou un autre correcteur a noté chaque tâche.

Les correctrices et correcteurs ont reçu une formation pour utiliser les grilles de notation uniques pour chaque tâche. Chaque grille contenait des descripteurs de travaux d'élèves pour tous les niveaux du continuum de notation. Des exemples de travaux d'élèves représentant chaque point sur ce continuum ont été utilisés avec les grilles. Pour chacune des quatre tâches d'écriture (résumé, texte d'opinion en quelques paragraphes, nouvelle journalistique, texte informatif) une note de 0, 1, 2, 3 ou 4 a été attribuée à l'aide d'une grille de notation selon l'approche globale.

Notation selon l'approche globale, Phase 2. Les travaux d'élèves qui n'ont pas rempli les critères de réussite durant la Phase 1 ont été notés une seconde fois selon l'approche globale. Si le travail atteignait le seuil de réussite durant la Phase 2, l'élève avait réussi la composante écriture.

Pour chaque phase, quatre correctrices et correcteurs différents ont noté le travail de la ou du même élève, selon l'approche globale, produisant quatre notes obtenues selon l'approche globale, une pour chacune des quatre tâches d'écriture. La note totale obtenue selon l'approche globale a été déterminée en combinant les notes individuelles pour chaque tâche.

Notation analytique. Les travaux d'élèves qui n'avaient pas atteint le seuil de réussite durant la Phase 2 de la notation selon l'approche globale ont été notés à nouveau à l'aide d'une grille de notation analytique. La notation analytique est effectuée afin de permettre de donner aux élèves plus de commentaires au sujet de leurs travaux, afin qu'ils puissent s'inscrire à des programmes d'appoint.

Établissement d'un lien entre le TPCL administré en février 2002 et le TPCL administré en octobre 2002 : Description du processus d'établissement des normes en écriture
L'exercice d'établissement des normes, qui a fait appel à une méthodologie de classification des cahiers, a eu lieu les 19 et 20 novembre 2002, au centre de notation de l'OQRE. Onze personnes ont participé à l'activité, soit cinq expertes et experts de langue française et six de langue anglaise. Les onze expertes et experts provenaient de différents milieux, incluant des expertes et experts de l'OQRE dans les matières, du personnel d'universités, des représentantes et représentants du secteur privé et du personnel du ministère.

Voici une description des étapes suivies durant le processus d'établissement des normes :

Première étape : formation des membres du panel
Une équipe de quatre personnes composée de trois agentes et agents d'éducation de l'OQRE qui avaient participé à l'élaboration du test et des grilles de notation, et d'une personne spécialisée en psychométrie de l'Université de Toronto, ont donné la formation aux membres du panel. La formation donnée à l'équipe comprenait l'occasion de se familiariser complètement avec le contenu et les attentes du test, les grilles de notation des tests de février 2002 et d'octobre 2002, et des échantillons de travaux d'élèves procurant des exemples de rendements de tâches d'écriture différents. Ils ont également noté des travaux d'élèves à l'aide de la grille d'évaluation.

Deuxième étape : création de groupes
Les membres du panel ont été répartis en quatre équipes – deux équipes francophones et deux équipes anglophones. Les équipes francophones ont travaillé sur les cahiers de l'élève remplis en français, alors que l'équipe anglophone a travaillé sur les cahiers de l'élève remplis en anglais. On souhaitait de cette façon minimiser l'interaction entre les équipes.

Troisième étape : examen du matériel du test de février 2002
On a remis à chaque équipe 30 ensembles de cahiers de l'élève provenant de l'administration de février 2002, mettant en séquence les notes totales selon l'approche globale à partir des plus basses jusqu'aux plus élevées, les notes et les décisions de réussite ou de non réussite indiquées. Chaque membre de l'équipe a lu les quatre réponses dans chaque ensemble et a décidé si l'ensemble au complet méritait « une réussite certaine », « une chance raisonnable de réussite », « une chance raisonnable de non réussite » ou « une non réussite certaine ». Chaque membre de l'équipe a fait ce travail indépendamment sans en discuter avec les autres membres de l'équipe. Les membres de l'équipe ont ensuite indiqué leur décision sur des feuilles de données sur lesquelles paraissait leur nom. Ensuite, les membres de l'équipe ont discuté de leurs décisions individuelles et pris une décision en groupe pour chaque ensemble. Chaque groupe a également indiqué la décision obtenue par consensus sur la feuille de données sur laquelle le nom de son équipe paraissait. Ce processus a été répété avec 30 autres ensembles de cahiers de l'élève par équipe, bien que cette fois, les cahiers de l'élève n'aient pas été mis en ordre ou présentés avec des notes ou des décisions de réussite ou de non réussite.

Quatrième étape : examen du matériel du test d'octobre 2002

Tous les éléments du processus utilisé pour le travail fait pour l'administration de février 2002 ont été répétés pour l'administration d'octobre 2002. Les résultats de toutes les étapes du processus de prise de décision ont été présentés.

Cinquième étape : comparaison du matériel des tests de février 2002 et d'octobre 2002
Quinze ensembles assemblés en paires selon la note totale obtenue en fonction de l'approche globale ont été présentés aux équipes. Les équipes ont décidé si chaque ensemble représentait : « une réussite certaine », une « chance raisonnable de réussir », une « chance raisonnable de ne pas réussir » ou « une non réussite certaine ». De plus, elles ont indiqué quel élément de chaque paire était le plus fort. Toutes les décisions ont été présentées.

Résultats pour la composante écriture
Le personnel de l'OQRE, en consultation avec des expertes et experts en évaluation de l'extérieur, ont discuté des résultats du processus d'établissement des normes. Comme nous l'avons déjà mentionné, pour chacune des quatre tâches d'écriture (résumé, texte d'opinion en quelques paragraphes, nouvelle journalistique, texte informatif), une note de 0, 1, 2, 3 ou 4 a été attribuée à l'aide d'une grille de notation selon l'approche globale. Ces notes ont ensuite été pondérées afin de les faire concorder avec le processus de mise en équivalence. Les pondérations étaient les suivantes : 0 = 0, 1 = 10, 2 = 20, 3 = 35 et 4 = 45. Cette pondération différentielle permet d'assurer que seuls les élèves dont le rendement est uniforme aux quatre tâches d'écriture réussissent la composante écriture du test.

Un seuil de réussite final pour la composante écriture a été établi à 100 points sur un maximum de 180 points.

Bibliographie

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, New Jersey: Educational Testing Service.

Assessment Systems Corporation. *XCALIBRE User's Manual.* http://www.assess.com.

Barnard, J. J. (September, 1996). In search for equity in educational measurement: Traditional versus modern equating methods. Paper presented at Association for the Study of Evaluation in Education in South Africa's national conference at the Human Sciences Research Council Conference Centre, Pretoria, South Africa.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Toronto: Holt, Rinehart and Winston, Inc.

Clustan. What is cluster analysis? Retrieved March 25, 2003, from the World Wide Web: http://www.clustan.com/what_is_cluster_analysis.html.

Kolen, M. J. & Brennan, R. L. (1999). *Test Equating: Methods and Practices.* New York: Springer-Verlag.

Muraki, E., Himbo, C. M., & Lee,Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*(4), 325–337.

National Association of State Boards of Education (1999). Setting cut scores on large-scale assessment. *The State Board Connection: Issues in Brief, 19*(2), 1–12.

Office de la qualité et de la responsabilité en éducation (2002). *Guide pour l'administration du Test provincial de compétences linguistiques, octobre 2002.* Toronto (Ontario) Imprimeur de la Reine pour l'Ontario.

—(2001). *Guide pour l'administration du Test provincial de compétences linguistiques, octobre 2001.* Toronto (Ontario) Imprimeur de la Reine pour l'Ontario.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. H. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 339–369). Mahwah, NJ: Lawrence Erlbaum.

Smisko, A., Twing, J. S. & Denny, P. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education, 13*(4), 333–342.

Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications.* Thousand Oaks, CA: Sage Publications.

Yang, W. & Houang, R. T. (1996). *The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using an anchor item design.* Paper presented at the Annual Meeting of the American Research Association, New York, NY.